AD________________

Award Number: W81XWH-06-1-0641

TITLE:
Detection of Tumor Suppressor Gene Mutations on 17p in DCIS

PRINCIPAL INVESTIGATOR:
Lesleyann Hawthorn, Ph.D

CONTRACTING ORGANIZATION:
Health Research Inc

Buffalo, NY 14263

REPORT DATE: August 2008

TYPE OF REPORT:
Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 01/08/2008	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 31 Aug 2006 – 31 Jul 2008

4. TITLE AND SUBTITLE
Detection of Tumor Suppressor Gene Mutations on 17p

in DCIS

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-06-1-0641

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Lesleyann Hawthorn, PhD

Email: LHAWTHORN@mail.mcg.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Health Research, Inc.

Buffalo, NY 14263

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The most powerful indicator of the location of TSGs in sporadic breast tumors has come from LOH studies. The implication is that a recessive mutation in the gene is "exposed" because the normal gene has been lost. DCIS is considered a precursor lesion of infiltrating ductal carcinoma (IDC). LOH at 17p is a recurrent observation specific to grade III DCIS and Grade III IDC suggesting a role for these alterations in tumor progression. Since loss of 17p is categorically related to both high grade DCIS and high grade IDC this region more than likely harbors one or more tumor suppressor genes involved in the progression of DCIS to IDC. High-density oligonucleotide arrays offer the ability to sequence large numbers of loci in parallel using an automated approach. There are many examples where array-based sequencing has proved successful, however , most of these applications have used normal samples for the identification of SNPs in specific chromosomal regions. The CustomSeq Arrays enable the analysis of 300kb stranded sequence on a single array. This provides the most cost effective and efficient scheme to query large amounts of sequence in a single experiment. . Our plan is to use this technology to search for mutations at 17p13 in IDC cells that display loss of this region, suggesting the presence of mutated TSGs.

15. SUBJECT TERMS
Ductal carcinoma in situ, infiltrating ductal carcinoma, chromosome 17p sequencing by hybridization

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	9	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION

The most powerful indicator of the location of TSGs in sporadic breast tumors has come from loss of heterozygosity (LOH) studies. The implication here is that a recessive mutation in the critical gene is "exposed" because the normal gene has been lost. Despite these overwhelming conclusions, identifying the cancer genes in these particular regions has been a generally unfruitful endeavor. This lack of success stems from the extensive regions involved and the need to characterize large numbers of genes on a single gene basis.

Ductal carcinoma of the breast (DCIS) is considered a precursor lesion of infiltrating ductal carcinoma (IDC). The molecular events that trigger DCIS to progress to IDC are not understood for the most part, however, increasing evidence suggests that specific deletions of chromosomal regions may be significant. LOH at 17p is a recurrent observation specific to grade III DCIS and Grade III IDC suggesting a role for these alterations in tumor progression. Since loss of 17p is categorically related to both high grade DCIS and high grade IDC this region more than likely harbors one or more tumor suppressor genes involved in the progression of DCIS to IDC.

High-density oligonucleotide arrays offer the ability to sequence large numbers of loci in parallel using an automated approach. There are many examples where array-based sequencing has proved successful, however, most of these applications have used normal samples for the identification of SNPs in specific chromosomal regions. Our plan is to use this technology to search for mutations at 17p13 in IDC cells that display loss of this region, suggesting the presence of mutated TSGs. The CustomSeq Arrays enable the analysis of 300kb stranded sequence on a single array. This provides the most cost effective and efficient scheme to query large amounts of sequence in a single experiment.

The following lists the tasks we planned to complete within the year allotted for the research.

1: Identify IDCs with 17p13 loss using aCGH

We will use SNP-array CGH to identify IDCs with LOH at 17p13. We have already identified 8 such tumors and one cell line using this approach.

2. Design and validate primer pairs Primer selection and validation prior to commissioning the production of the sequencing arrays is essential in order to ensure that the array contains only those sequences for which optimal PCR products can be obtained. We had engineered a program to systematically design the primers to amplify all of the 1389 exons that map to 17p13 including the p53 gene. The validation of the 1200 PCR products proved to be problematic. We had to spend an inordinate amount of time on the validation including running gels for size determination and running spectrophotomer analyses to determine the concentration of the products. There appeared to be a high failure rate of the PCR reactions. Since that time new technologies have become available to enrich for DNA sequences and we are now in the process of developing this technology

3. Design array and perform hy bridizations We have designed and commissioned the arrays and now have 45 of these awaiting hybridization.

4. Evaluate identified mutations in 300 primary breast tumors

We were prevented from completing all the tasks listed within the allotted time due to unforeseen circumstances. The major impediment to our research arose due to a restructuring of the Pathology dept at RPCI. We were unable to obtain any tissues until late June of 2007. We did, however make inroads with the rest of our design and plan to continue these experiments until they are completed. No inroads were made toward achieving Aim 4 as this depended on data accumulated from the other 3 Specific Aims. The 17p CustomSeq array has been designed and commissioned. We have used SNP arrays to demonstrate the capability of this technology to identify tumors carrying a loss of chromosome 17p. We have designed and ordered the primer pairs to amplify all the known coding sequences on 17p. We have expended considerable effort in working out the optimal conditions for PCR cleanup and amplicon quantification using a breast cancer cell line that has a demonstrated loss of 17p. Since that time we have revised our approach to move away from PCR-based amplification of the exons and are currently developing an array-based DNA sequence capture approach which will alleviate the problems we were facing

BODY

For **Specific Aim 1** we had planned to identify IDCs with 17p13 loss using aCGH. We opted to use SNP arrays as for the identification of these tumors since it would also provide simultaneous LOH and copy number information on other chromosome which may have impact on the clinical outcomes. Since we only recently received the tumors, the main focus has been to obtain excellent quality DNA. Most of the samples we have been receiving are either very fatty or very fibrous. A number of DNA isolation procedures have been evaluated for these samples. The results have shown that attempting to obtain DNA and RNA from the same protocol (Trizol and Quiagen Lipid Tissue Extraction kit) do not provide quality DNA. The highest quality DNA was obtained from using the standard Phenol/Chlorophorm extraction. This is now the protocol that will be used to isolate all the DNA from the tumor samples.

We also performed CGH analysis on the breast cancer cell line MDA-MB-468 using the 250K SNP mapping array. This analysis gave us a high resolution profile of this cell line and confirmed that the line had retained a single copy loss of chromosome 17 (see figure 1).

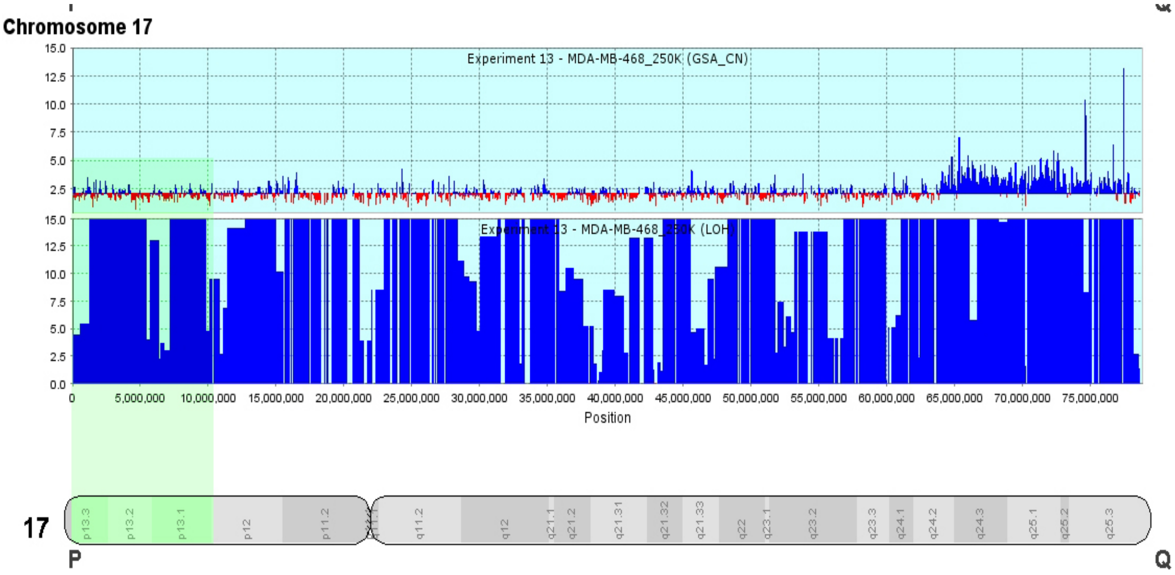

Figure 1: Copy number and LOH data from the breast cancer cell line MDA MB-468 using the 250K mapping array (Affymetrix). All of chromosome 17 is demonstrating LOH. The region that we will sequence using our customSeq array is highlighted in green. It can be seen that this region is showing copy number loss (shown in red) as well as LOH (bottom graph). This is in contrast to the telomeric region of the Q arm, where LOH is increased, with an accompanying amplification. This data was graphed using Exemplar Copy Number Tool (Sapio Biosciences)

This cell line provided the ideal tool for the troubleshooting phase of the protocols. We could use this cell line to test the primers and arrays as described in **Specific Aim 2.** Our first strategy was to employ a multiplex approach for the reasons of limiting the amount of reagents and samples required. For this approach we designed the primer sets in groups of 10 that had similar annealing temperatures and amplified smaller fragments usually encompassing 1-2 individual exons and ~1000 bp maximal in size. A second requirement was that each amplicon had to differ enough in size to allow discrimination using gel electrophoresis. Figure 2 shows a typical gel from some of these PCR reactions as can be seen in some instances the larger products failed amplify optimally. An additional concern at this stage was that we needed to add equimolar amounts of reactions to the final hybridization cocktail and the multiplex reactions would not allow this.

When the time came to design primers for the remaining amplicons to hybridize to the rest of the array, we reconfigured our PCR primer strategy to incorporate larger fragments. Using a longer range approach, primers were designed to amplify 371 individual fragments ranging in size form

300 to 6800bp. We attempted to validate the amplification products but again it was costly and very time consuming. Recently, developed **Target enrichment,** also known as genome partitioning, is a method that isolates specific fragments of genomic DNA for sequencing. A library of complementary oligonucleotide "baits" is constructed and used to harvest fragments of interest (target DNA). The target DNA hybridizes well with the baits, but other DNA does not, which forms the basis of a powerful selection method.. Agilent eARRAY has predesigned probes for exons and splice variants of all RefSeq genes. For most applications we will use these pre-validated probes for the potential TSGs of interest. In situations where probes may not be available we will design custom probes we will use the design software implemented in eArray. Cross-hybridization between exons is usually not extensive, except in those cases where specific gene family members share highly homologous sequences and are located in the same genomic region. For each probe designed we will analyze the probe sequence for repetitive elements using REPEAT MASKER and for homologous sequences using BLAST. For each Agilent library, 55K of unique 120mer oligos are synthesized on a wafer, the oligos are then eluted and *in vitro* translated into biotinylated mRNAs. The library from Agilent is received as a kit that contains a set of biotinylated RNA oligonucleotides. However, as part of the library manufacturing process, Agilent first creates DNA oligonucleotides, and then later transcribes them into RNA. We have opted to use this approach (as opposed to Nimblegen) because it is less expensive and the amount of starting DNA required (1-5 ug vs 750ng) is much less and more amenable to DNA extraction from FFPE samples.

Specific Aim 3 involves the design of the array and this section of the proposal has been completedThe final product contained representation of 142 genes. This amounted to 1389 individual exons with a total length of 299868 bases to be sequenced. By combining smaller exons in our primer design we were able to reduce the number of PCR reactions to a manageable 371 reactions. In order to determine if we could detect deletions we added 3 additional genomic regions on the array that had known deletions in samples that would be spiked in for those specific sequences.

Target coding sequences and intron-exon boundary regions critical for RNA splicing were extracted. We obtained the most updated human genome sequences and the RefGene annotation files from the UCSC genome website at http://genome.ucsc.edu. For the intron-exon boundary region, we include 20 bp at the 5' end of each exon and 6 bp at the 3' end of exon. For alternatively spliced genes, the overlapping regions of all alternative forms were extracted to insure that all possible coding regions were represented on the array. For genes that were located at the same genomic region but on different strands, only the union of the coding sequences on one strand was used in the design in order to preserve "real-estate" on the array and to prevent the target concentration from being diluted by competing probes on the array. The probes for each target base are sequenced from both directions, so none of the genes will be missed. This step was implemented as a Perl program (**getCDS4chip.pl**). The program takes a refGene annotation file and a chromosome sequence file as the inputs and outputs the resulting CDS for a selected chromosome or a region in fasta format with each CDS region as a separate entry.

In order to omit cross-hybridization reactions among target CDS sequences, the extracted target sequences are submitted to Affymetrix chip design group to double-check the sequences for potential cross-hybridization events. Cross-hybridization between exons is usually not very extensive except in those cases where a specific gene family shares highly homologous sequences and are located in the same genomic region. We only considered the exact 25 bp probe sequence match as cross hybridization events, with the 13th base being ignored since the re-sequencing chip technology relies on its ability to distinguish the perfect match (PM) and mismatch (MM) sequences. When we encountered the exact same probe design for two different exons, one of the target bases was masked using Affymetrix chip design instruction file. This occurred in instances where the number of identical probes exceeded a selected cutoff value. We used a "greedy method" to mask bases with identical probes by masking the exons with maximal number of identical probes with other exons first. Two programs were developed for this step. The first program, **SumSimilarProbes.pl**, takes the target CDS sequence file and the Affymetrix similar probe file as input and outputs a summary of identical probes found for each region. Another program, **MaskSimilarProbe.pl**, takes the target CDS input file and output of the previous program as input and outputs a chip design instruction file with masked probes.

Control sequences

Aside from the normal B2 sequences on the array that allow for grid alignment, and act as hybridization controls, a plasmid sequence is included to IQEX is included to control for amplification and labeling. Our array design also included sequences that would allow us to determine if small deletions could be detected using this array-based sequencing approach. In order to achieve this goal, sequences from the MAP1S, NCOR1, B2M were tiled on the array. These genes that have been demonstrated to have small deletions in the colon cancer cell lines LS180, LoVo and RKO as determined by our group using a nonsense mediated decay (NMD) approach (Ivanov et al 2007). Table X below shows which cell lines contain the deletions. Sequences from these genes will be amplified and spiked into the hybridization cocktail. If the probes representing these sequences tiled on the array do not show intensity at these specific nucleotides then small deletions are detectable using this approach.

GENE NAME	GENE SYMBOL	MUTATION	CELL LINE
microtubule-associated protein 1S	MAP1S	Deletion $(AG)_4$ nt 1447-1454	LS180
beta-2 microglobulin	B2M	Deletion $(CT)_4$ nt 37-44	LoVo
Nuclear receptor co-repressor 1	NCOR1	Deletion $(TG)_2$ nt2463-2466	RKO

Table 1: Spike in control for detection of microdeletions.

The results of this rigorous design strategy was a continuous ordered list of exon regions although in a very limited number of cases homologous regions of exons are masked out.

Hybridization Optimization

A large amount of effort was put into finding the optimal hybridization conditions. Several “spike in” experiments were performed on the arrays to compare a defined set of sequences. It was necessary to define the optimal experimental parameters since these types of experiments had not previously been attempted.

For the initial experiments we ran small range products amplified from a MDA-MB 467 cell line. The pooled hybridization cocktail contained probes that would only hybridize to half of the array. Figure 3 shows that nonspecific hybridization was not observed, and no base calls were assigned to the sequence corresponding to the absent fragments. These results show a close to a complete absence of nonspecific hybridization.

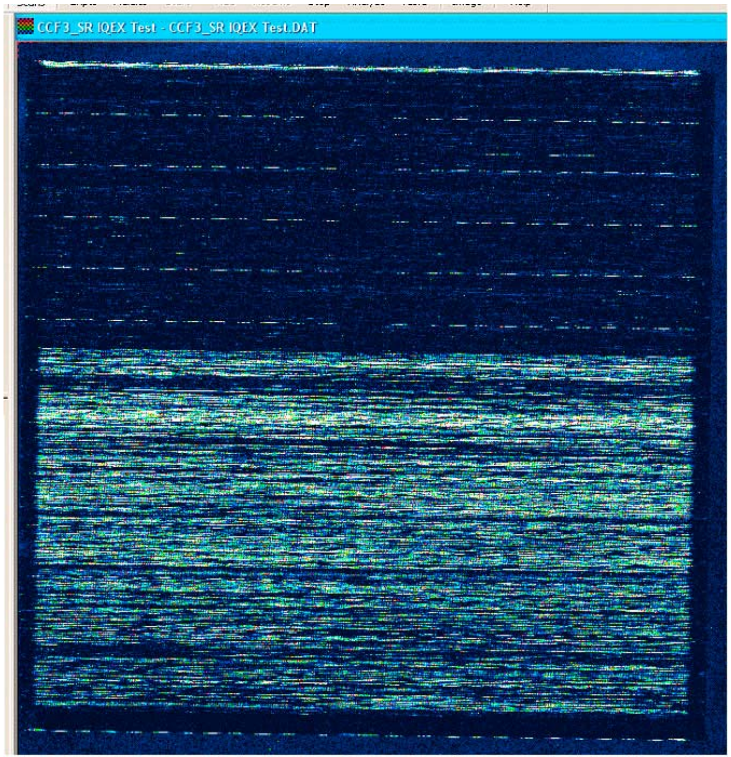

Figure 3: CustomSeq Array for Chromosome 17p Demonstrating Specific Hybridization: The hybridization cocktail contained “spike in” PCR products that would bind to sequences covering roughly half of the array. The bands visible in the upper half of the array are B2 grid alignment sequences. The illuminated band across the top of the array is a positive control plasmid sequence IQEX that is spiked into the hybridization cocktail. This experiment demonstrate the specificity of the PCR products for the sequences tiled on the array.

The results of this initial experiment were very encouraging. The PCR pool contained 476 indivdual exon products (short range PCR amplicons). Of these 377 were successfully sequenced. Of the 99 exons that failed on the array, 82 were weak or absent PCR products when analyzed using gel electrophoresis. Another 17 had passed at the PCR level but had achieved less than 80% quality score on the array. “ No calls” are assigned to a base when the algorithm could not call the base. NO CALLS typically result from either saturation of the hybridization signal, large signal to noise ratio, or weak signals. We have found that the majority of no calls were G or C and these reside in GC rich regions. These quality scores would be expected to be improved with additional samples added to the analysis. In summary, once the failed PCR reactions were subtracted from the analysis we had achieved a 95.5% call rate on our first analysis.

Strategy to identify known SNPs and identify non-synonymous SNPs

The re-sequencing results will generate thousands of SNPs, most of them as being known and synonymous. We need to filter these SNPs out and narrow down to a smaller set of interesting SNPs. We have developed the following strategy for this purpose.

1. Compute the genomic position for each SNP call generated in our experiments: The data generated by the chip are SNP calls for each position relative to the individual sequence target regions. Since we have recorded the genomic positions of each target in the sequence file, it is very straightforward to convert the position of the SNPs into the universal genome coordinates.
2. Identify known SNPs. The known SNP position can be downloaded from UCSC genome center **(snp.126.txt.gz).** By sorting and comparing the known SNP positions with our SNP set, we can identify known SNPs in our results easily.
3. Identify non-synonymous SNPs. For each refSeq, the coding region start position is known, so it’s easy to identify the new codon for each SNP. Non-synonymous SNPs can be detected easily by comparing the corresponding amino acids for both new and original codon can identify. Some of the refSeq is not fully annotated, we can use the protein sequence downloaded from NCBI to identify the reading frame. The detected non- synonymous occurring at significant frequency will be selected for verification in the web lab.

This strategy was implemented as a perl program called **analyzeSNP.pl.** It takes three files: the CDS sequence file, the resulting SNP file generated by the Affymetrix software, and the known SNP file, as inputs. The program outputs the classification of each SNP as known synonymous, known non-synonymous, unknown synonymous, and unknown non-synonymous.

Key Accomplishments

1. Unique sequence primers designed to amplify all coding sequences on 17p13.
2. CUstomSeq array designed to sequence all coding sequences on 17p13.
3. Protocol for sequenencing- by – hybridization streamlined.
4. Developed crucial softwares that will be used for the design of primers and the array.
   in-house modified version of Primer3
   UCSC genome browser (in-house version)
   ChipCDS2Primer.pl.
   getCDS4chip.pl
   SumSimilarProbes.pl
   MaskSimilarProbe.pl

5. Developed software for the analysis of sequences.
   snp.126.txt.gz
   analyzeSNP.pl.

Reportable Outcomes

N/A

Conclusions

This has described our progress to date on this project. We feel we have obtained some very useful knowledge, developed some key programs and have proven that our approach is a valid one. We have again revised our strategy so that we will use array-based sequence capture to obtain the exonic sequences from the breast tumors. This approach will alleviate the obstruction that was presented by the PCR-bases approach. We are looking forward to using this new approach to continue with this important work.